Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer

Rabinovici-Cohen, Simona; Tlusty, Tal; Abutbul, Ami; Antila, Kari; Fernandez, Xosé; Grandal Rejo, Beatriz; Hexter, Efrat; Hijano Cubelos, Oliver; Khateeb, Abed; Pajula, Juha; Perek, Shaked

Published in:
Medical Imaging 2020

DOI:
10.1117/12.2551374

Published: 01/01/2020

Document Version
Publisher's final version

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer

Rabinovici-Cohen, Simona, Tlusty, Tal, Abutbul, Ami, Antila, Kari, Fernandez, Xosé, et al.

Simona Rabinovici-Cohen, Tal Tlusty, Ami Abutbul, Kari Antila, Xosé Fernandez, Beatriz Grandal Rejo, Efrat Hexter, Oliver Hijano Cubelos, Abed Khateeb, Juha Pajula, Shaked Perek, "Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer," Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 113181B (2 March 2020); doi: 10.1117/12.2551374

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer

Simona Rabinovici-Cohen[a], Tal Tlusty[a], Ami Abutbul[a], Kari Antila[c], Xosé Fernandez[b], Beatriz Grandal Rejo[b], Efrat Hexter[a], Oliver Hijano Cubelos[b], Abed Khateeb[a], Juha Pajula[c], Shaked Perek[a]
[a]IBM Research – Haifa, Mount Carmel, Haifa 3498825, Israel
[b]Institut Curie, 26 Rue d'Ulm, 75005 Paris, France
[c]VTT Technical Research Centre, Vuorimiehentie 3, Espoo, Finland

ABSTRACT

Women who are diagnosed with breast cancer are referred to Neoadjuvant Chemotherapy Treatment (NACT) before surgery when treatment guidelines indicate that. Achieving complete response in this treatment is correlated with improved overall survival compared with those experiencing a partial or no response at all. In this paper, we explore multi modal clinical and radiomics metrics including quantitative features from medical imaging, to assess in advance complete response to NACT. Our dataset consists of a cohort from Institut Curie with 1383 patients; from which 528 patients have mammogram imaging. We analyze the data via image processing, machine learning and deep learning algorithms to increase the set of discriminating features and create effective models. Our results show ability to classify the data in this problem settings, using the clinical data. We then show the possible improvement we may achieve in combining clinical and mammogram data measured by the AUC, sensitivity and specificity. We show that for our cohort the overall model achieves sensitivity 0.954 while keeping good specificity of 0.222. This means that almost all patients that achieved pathologic complete response will also be correctly classified by our model. At the same time, for 22% of the patients, the model could correctly predict in advance that they won't achieve pathologic complete response, enabling them to reassess in advance this treatment. We also describe our system architecture that includes the Biomedical Framework, a platform to create configurable reusable pipelines and expose them as micro-services on-premise or in-the-cloud.

Keywords: Neoadjuvant Chemotherapy Treatment, Neoadjuvant Treatment, Breast Cancer, Radiomics, Big Data, Artificial Intelligence, Deep Learning, Machine Learning

1. INTRODUCTION

Neoadjuvant Treatment (NAT) in which systemic treatment of breast cancer is administered prior to surgical therapy, is one of the approaches in breast cancer treatment. These can include chemotherapy, radiation therapy, hormone therapy, targeted therapy or immunotherapy which are given before surgery. Our study is mostly on patients that received Neoadjuvant Chemotherapy Treatment (NACT). Potential clinical advantages of NACT have been largely studied and include improved overall survival, improving the rate of breast-conserving therapy and obtaining accurate in vivo tumor sensitivity. Initially this treatment was employed for patients with inoperable disease, but over past years it proved beneficial in other patients including those with early-stage, operable breast cancer[1]. A preferred outcome[2,3] of this treatment is pathological complete response (pCR) namely absence of invasive residual tumor in the breast, and of invasive disease in the axillary nodes. Achieving a pathologic complete response after neoadjuvant chemotherapy is correlated with improved disease-free survival and overall survival compared with those experiencing a partial or no response to neoadjuvant chemotherapy. Today, the clinical parameters to select the NACT option is based on breast cancer subtype, tumor size, disease grade, number of affected nodes, age and tumor growth. Imaging is used to evaluate the position of the tumor and its size. However, quantitative models based on clinical and imaging features are not considered.

Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications,
edited by Po-Hao Chen, Thomas M. Deserno, Proc. of SPIE Vol. 11318, 113181B
 · doi: 10.1117/12.2551374

Radiomics is an emerging area of intensive interest in the medical community[4]. It aims to extract large number of quantitative features from multi modal medical images, and thus it is an important enablement of precision medicine. It has the potential to improve the prediction of prognosis and therapeutic response to medical conditions because of two new advances. First, there is a continuous growth in the number of medical images and in the amount of information expressed in each image (better resolution), providing us with a large dataset of individual patient information. Second, the field of computer vision is rapidly progressing, incorporating new advanced big data algorithms from image processing, machine learning (ML), deep neural networks learning (DL) and general artificial intelligence (AI) techniques.

In this paper, we describe a system that utilizes radiomics and provides results of a pilot to predict breast cancer NACT pathologic complete response on a cohort from Institut Curie. The system employs a scalable and generic architecture for big data that can be deployed on-premise or on-the-cloud and is specifically designed for analyzing multi-modal data with artificial intelligence. The system is being developed as a part of the breast cancer pilot of the BigMedilytics project[5], which aims to use big data technologies to achieve productivity in the healthcare and life sciences sector by reducing costs, improving patient outcome and delivering better access to healthcare facilities.

Our main contribution in this paper is the architecture of a radiomics system and the methods to improve NACT response prediction using multi-modal data and of different types. We describe an ML model for clinical data, a DL model for mammogram (MG) data, and an ensemble model of the individual clinical and MG models. To our knowledge, this is the first time a DL model with texture extractions is used on mammograms for NACT prediction. We evaluate our models on a cohort from Institut Curie with over 1700 patients, out of which 528 patients have MG imaging and show the results to discriminate complete responders versus partial or none responders' patients. Additionally, we report on data challenges with MRI modality when the data is not collected for academic research purposes.

We describe the work related to this topic in section 2, the methods used to develop our multi-model predictor in section 3, the big data prediction system employed for the pilot in section 4, the results of our models in section 5, and discuss these results in section 6. Finally, we conclude our work and describe future improvements in section 7.

2. RELATED WORK

Predicting NAC therapy response has been previously explored via various methods, but most of these methods used a single modality, and only a few methods combined clinical and imaging to improve the prediction. Parekh et al.[6] defined a multiparametric imaging radiomic framework termed MPRAD for extraction of radiomic features from MRI images. They use a volume created from proton density (PD) T2 weighted, T1 weighted, diffusion weighted (DWI) and perfusion weighted imaging (PWI) and extract a gray level vector for each pixel used for statistical features based on first and second order analysis. In another paper by Scheel et al.[7], the authors compared longest diameter of tumor information from different imaging modalities (MRI, MG and clinical) and assessed which best correlated with a response prediction to NACT. They found that MRI was the best indicator, more than a combination of modalities.

Using mammography to predict pCR response is not widely used. Two papers in the field used indicators whether information is present or not in MG to predict a positive NACT response. Savaridas et al.[8] extracted such features from MG and US images. In US it was the presence of posterior shadowing and in MG they examined presence of micro-calcifications and tumor speculations. They also examined the correlation of tumor size with the prediction, as this relationship was demonstrated by Ring et al.[9] as well. Ring did not combine clinical features; but they did check the imaging indicators in subgroups according to tumor subtype (HER2 positive and triple negative). They found correlation but only for HER2, while for triple negative no features were correlated with the prediction. Another paper by Li et al.[10] examined this relationship but used the presence of calcifications in MG images. They found that calcifications were not changed after NACT and concluded there is no relation between MG and pCR.

A combination of clinical and textural imaging features using a logistic regression was performed by Lee et al[11]. The authors used clinical and pathological features composed from: sex, age at initial diagnosis, body mass index (BMI), clinical TNM staging, estrogen receptor (ER), progesterone receptor (PR) expression, HER2 expression, histological grade and Ki-67 expression. Co-occurrence matrix, histogram and gradient based algorithms were used on PET/CT images. This combined model scored significantly higher than only clinicopathological factors model.

A retrospective study regarding clinical factors that are important in the prediction of a tumor response to NACT was conducted by Fasching et al.[12] They found that Ki-67 is an independent predictor for pCR in all patients across all subtypes. They also affirmed findings from previous publications that parameters such as age, BMI, hormone receptors and more are correlated with pCR.

Using deep learning to predict NACT response was firstly done by Ravichandran et al.[13] in the context of DCE-MRI scans and an open NACT dataset of 166 patients. A CNN utilizing a combination of both pre- and post-contrast images best distinguished responders. To that they added prognostic clinical variables (age, largest diameter, hormone receptor and HER2 status) and found that inclusion of HER2 status could further improve capability to predict NACT response. In our study we investigate the use of clinical data with MG imaging to the same problem setting. MG scans are generally acquired via a standardized protocol, are routinely taken for NACT patients and proved to include informative features in screening use cases; thus interesting to investigate for our use case as well.

In our work we demonstrate pCR prediction from clinical data, including factors that are previously reported as correlated with pCR and show their consistent importance in our dataset as well. We also extract features from MG and examine its contribution with clinical information to predict NACT response. We show a system to deploy pipelines of trained models in heterogenous environments. We show results regarding our different experiments and how sensitivity and specificity characteristics are chosen.

3. METHODS

3.1 Overview

Our overall method is an ensemble of several models applied on the various modalities: ML models are applied on the clinical data, and a pre trained DL model followed by texture feature extraction is applied to the MG data. As only about half of the patients have MG data, our ensemble method takes in account the case of partial data. In addition to the selected models, we describe an experimental simple model on MRI data which did not result in a meaningful prediction, and hence was not added to the final ensemble.

3.2 Data preparation

The Curie dataset includes a cohort of 1738 patients that received NACT between 2012 and 2018. Out of them, 1383 patients had binary classification regarding the complete response to treatment that substitutes our ground truth. This cohort was identified and extracted from Curie expert clinical repositories. We developed a pipeline to create the dataset that queried, curated, merged and anonymized heterogeneous data from several data sources. Part of the data was already manually curated and structured in Curie MySQL repositories. Some items were extracted from the text using Natural Language Processing and Machine Learning approaches, then stored in an Elasticsearch database. Finally, additional pilot specific information as treatment response was extracted from the medical reports. Thus, the cohort includes (1) demographics such as age at diagnosis, weight and height; (2) tumor properties such as laterality, breast cancer histology, grade of the tumor, percentage of stromal tumor-infiltrating lymphocytes, hormonal receptors (estrogen, progesterone, HER2); (3) treatment characteristics such as chemotherapy protocol, properties of the surgery, radiation dose; (4) evolution after treatment such as relapse and metastatic events; (5) pCR classification. Table 1 provides some statistics on part of the data.

The clinical data protected health information (PHI) was fully anonymized. During the anonymization procedure of the clinical data we removed all the patient's identifiers such as patient name, address or date of birth. Multiple-choice variables (such as side of the tumor, histological type, type of breast cancer, etc.) were mapped to integers and the correspondence map was separated from the dataset. Finally, all the dates (date of diagnosis, chemotherapy, surgery, etc.) were replaced by a single integer, indicating the number of days elapsed since the date of birth. The resulting dataset thus contained only numerical values with no relation whatsoever to the patient identifiable data, which is both helpful to develop AI solutions and preserves the privacy of the patients. For the imaging dataset, each type of modality has a different anonymization procedure as they have some different DICOM tags. Some of these tags are nominative, like name or date of birth of the patient, hence were anonymized. A priori there is no need to modify the pixel data of the DICOM images. In addition, we excluded cases of multifocal and bilateral breast cancer, cases with skin invasive or inflammatory tumors, and cases who have relapsed from a previous tumor event.

Table 1: Curie clinical data cohort

	pCR	Non-pCR
Total	26.5%	73.5%
Histological type		
Ductal	28.5%	71.5%
Lobular	10.5%	89.5%
Nottingham grade		
I	26.0%	74.%%
II	20.0%	80.0%
III	33.0%	67.0%
Immunohistochemistry subtype		
TNBC	38.2%	61.8%
Luminal A	5.9%	94.1%
Luminal B	22.2%	77.8%
HER2+	39.0%	61.0%

3.3 Clinical Model

The 1383 patients that had ground truth were split to 5 equally distributed folds between positive and negative samples in the train and test splits (approximately 70% negative response and 30% positive response patients). We then use 26 pre-treatment features of the patient described in previous section, and train with three ML known algorithms, Random Forest, Logistic Regression, and Extreme Gradient Boosting[14] (XGBoost) for predicting NACT response. We perform cross validation and compute the receiver operating characteristic (ROC), area under the curve (AUC), sensitivity and specificity with confidence interval for each fold as well as the mean values. We also examine the features importance produced by our models, which seems to correlate with what is known from clinical studies.

3.4 MG Model

For MG, we utilize a DL model that was pretrained on IBM proprietary data, which consists of thousands of annotated mammograms to classify the existence of a tumor. This approach was selected as the number of MG images in Curie cohort is relatively small, 1261 images from 528 patients, which is not enough to train a deep learning network. Thus, we had to build on previous model that was trained on MG data to classify a different but related task. The model is a customized Inception-ResNet-V2 architecture by Szegedy[15]. It is composed of 14 Inception-Resent blocks and splits into two paths, one for global malignant prediction of the image and one for local tumor contour detection. This network is a variant of Le[16] where we share weights between paths, one used for classification of the global score and the second for producing a segmentation map. A gray scale input MG image and a binary mask created according to the tumor contour from radiologist annotations, are used as input to train the network and solve the task of classifying MG lesions to malignant lesion versus all other types (normal and benign). At inference time on Curie dataset, the network extracts a tumor malignancy prediction but also a heatmap which represents the tumor location. Figure 1 shows the output of the detection on an MG image and the tumor margins we use for feature extraction.

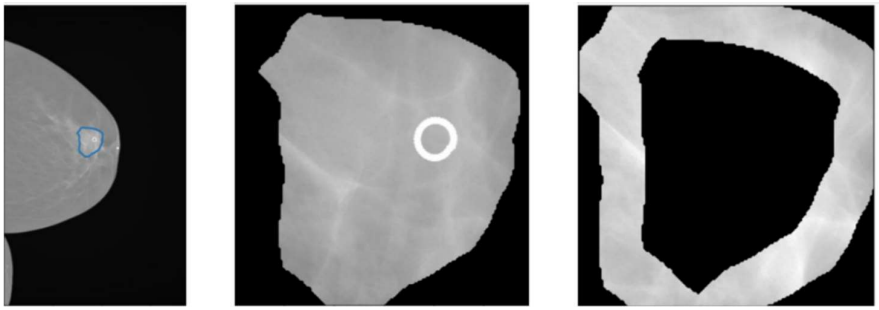

Figure 1. Network output predictions of tumor detection. Left: MG image from Curie dataset with a detected contour around tumor area. Middle: Tumor patch extracted from detected area. Right: Tumor margins extracted.

Detection of a tumor location with a binary mask enables us to extract texture features from the tumor area and the peritumoral area which is the margin of the tumor. A total of 30 radiomics texture features maps (24 Gabor, local binary pattern (LBP), Gray Level Co-occurrence Matrix (GLCM), 4 wavelets) and gray level distribution features were extracted pixelwise on each of the MG images. For each feature map, first order statistics (entropy, minimum, 10 percentile, 90 percentile, maximum, mean, median, standard deviation, skewness, kurtosis and mean absolute deviation) were computed within the tumor area and in the peritumoral area, summing up to 349 texture features per patient. We then train those features with Random Forest classifier performing cross validation to get the final MG score.

3.5 Experimental MRI Model

We use annotated MRI subtractions that have a rectangle around the tumor in a single axial slice. We segment the tumor by searching a bright intensity (defined by the 98th intensity percentile) from the annotated area, using that as a seed point and applying a simple 3D region growing algorithm. We then transfer the tumor mask from the subtraction to the baseline contrast image. From the masked area in the baseline contrast image a number of candidate features are extracted including the min/max/std of whole image grey values (minG/maxG/stdG), mean/min/max/std of segmentation grey values (meanf/minG/maxG/stdG), difference between segmentation and global (minG, maxG, stdG), x dimension length in mm of segmentation (xln), number of border voxels in segmentation (bordPix), and volume in cubic millimeters of border voxels (bordPix_norm). Feature weights are described in figure 2.

The individual contributions of these features in our classification task were tested with Logistic Regression, Support Vector Machines (SVM) and XGBoost. Additionally, separate feature selection test with chi-squared and ANOVA F-statistic was done. Finally, p-values on the chi-squared distribution were computed. The features with the significant p-values were meanG ($p < -10^5$) and bordPix_norm ($p = 0.029$). However, as this MRI model AUC for predicting NACT is low, it was not incorporated in the final ensemble model.

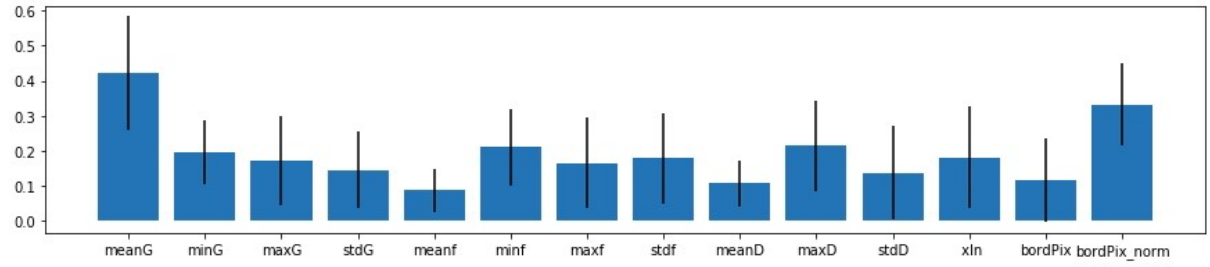

Figure 2. Feature weight coefficients for the best classifier model for stand-alone MR feature testing. Higher is better. The line segments above bars show standard deviation of the observed coefficients.

3.6 Ensemble Model

The ensemble model includes six models: three models based on clinical data and three models based on features extracted from the MG images. These models were trained with different seeds and classifiers (Random Forest, Logistic Regression, and XGBoost). Out of the 1383 patients that have clinical data, 528 patients have MG scans, where each patient may have several MG images. Thus, for patients with clinical and MG imaging we have six scores and for patients with only clinical data we have three scores from the clinical based models alone. Note that the score per patient of an MG model is a single score from the mean of the scores of all the patient images.

To ensemble the models, we examined several strategies and for each one of them evaluated the AUC, sensitivity and specificity. We first tried the stacking classifier in which we trained a meta model on top of the six models scores using same five folds as used for the individual models. We also tried several voting strategies, in which some of them consider the threshold of individual models. However, these models did not prove as effective as using a simple function of all individual scores, so the final selected model that provided the best result is the mean of all available scores per patient.

4. SYSTEM ARCHITECTURE

The methods described above come together in our designed system. The research and development cycle of ML and DL algorithms includes several steps that our system architecture needs to support. The cycle steps are: (1) Data preparation - collection, curation, anonymization, annotation of the ground truth, creation of data splits for training, test and held-out; (2) Algorithms development - training using various methods for the different modalities, test models utilizing relevant quantitative measures, select model or an ensemble of models; (3) Pipelines deployment - transforming the models to services, evaluate accuracy on held-out data split, deploy pipelines in scalable platform and run service on real-world data.

To support the above steps, our architecture needs to enable the integration of polyglot algorithms that may be written in different programming languages (e.g. Python, Java, C) and use different DL frameworks (e.g. TensorFlow, PyTorch). Moreover, the DL algorithms are computationally intensive and have special requirements that need to be considered and optimized such as GPU execution and large memory consumption. To support these requirements, we use the Biomedical Framework to create configurable reusable pipelines and expose them as microservices.

One important feature of the Biomedical Framework is the enablement of running pipelines of AI models on distributed environments such as SPARK cluster and doing that transparently to the algorithm developer. Given a description of the pipeline with the algorithms dependency graph, the Biomedical Framework automatically translates the pipeline descriptor to an efficient SPARK application. The resulted application is efficient in the sense that it minimizes the time from the beginning of the first algorithm until the completion of the final algorithm in the pipeline.

The overall pilot architecture is depicted in figure 3. To comply with regulations as GDPR, we use a model-to-data paradigm where all the data remains at Institut Curie infrastructure. All computations are applied on a strong GPU enabled server that resides in Curie, and various docker containers and pipelines of analytics models are transferred to the server and executed there. The overall flow is as follows: the anonymized imaging and clinical data are transferred from Curie expert repositories to the pilot server hosted within the institute infrastructure. Training and inference pipelines utilize the data and produce analytics results as HL7 FHIR[17] objects. The analytics results are stored in a repository and a viewer is used to visualize those analytics results and evaluate them against the patients ground truth.

This system executes the models we described in section 3. At inference time for a new patient, we can use this architecture and achieve a prediction alongside the patient's information, which will be depicted in a visual manner to the practitioner that will use the system.

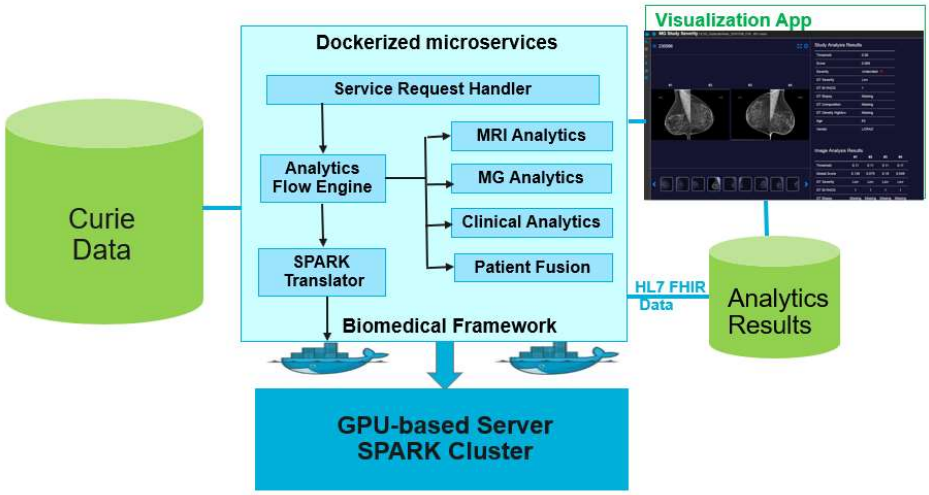

Figure 3. Pilot system architecture. On the left is the input data, middle part includes the pipelines micro-services and on the right are the stored results as HL7 FHIR objects and their visualization.

5. RESULTS

In the clinical data experiment, our results on the Curie dataset show promise regarding the ability to predict pCR using clinical parameters alone. In figure 4 left, we can see the ROC curve and respective AUC for the clinical only model using XGBoost classifier achieving 0.694 AUC (95% confidence interval [0.669, 0.718]). As our models use the pilot system architecture described in section 4, during training we set the sensitivity operation point to 0.99 and derive a threshold. This threshold is then used to compute sensitivity and specificity on the test data. With this approach, the sensitivity for clinical model on Curie test data is 0.882 and the specificity 0.326. The clinical models with Random Forest and Logistic Regression classifiers showed similar results, indicating that there is discriminating information in our dataset and this information is expressed in the clinical features without any dependency on a specific classifier. The important features found by the model are age, weight and height, KI67, HER2, estrogen status, progesterone status, with p-values of $p < 1e - 6$.

For MG, a naïve transfer learning of a DL model that classifies tumor existence, did not show much power by itself for classifying NACT response. However, using this model for segmentation on Curie data to detect tumor location and then extracting texture features from the patch detected by the model has provided useful information in combination with the clinical models.

The ensemble of clinical and MG models is depicted in figure 4 right, showing 0.708 AUC (95% confidence interval [0.683, 0.732]). As previously, we obtained the threshold for sensitivity 0.99 on the train data and used that threshold to compute sensitivity and specificity on the test data. With this approach, the ensemble model improves the sensitivity on Curie test data to 0.954 while keeping a good specificity score of 0.222.

We also compared the cross validation mean specificity of clinical only model versus the ensemble model at predefined sensitivity operation points. For sensitivity 0.99 we obtained specificity 0.112 in the clinical only model and specificity 0.118 in the ensemble model. For sensitivity 0.98 we obtained specificity 0.112 in the clinical only model and specificity 0.127 in the ensemble model. Finally, for sensitivity 0.95 we obtained specificity 0.261 in the clinical only model and specificity 0.265 in the ensemble model.

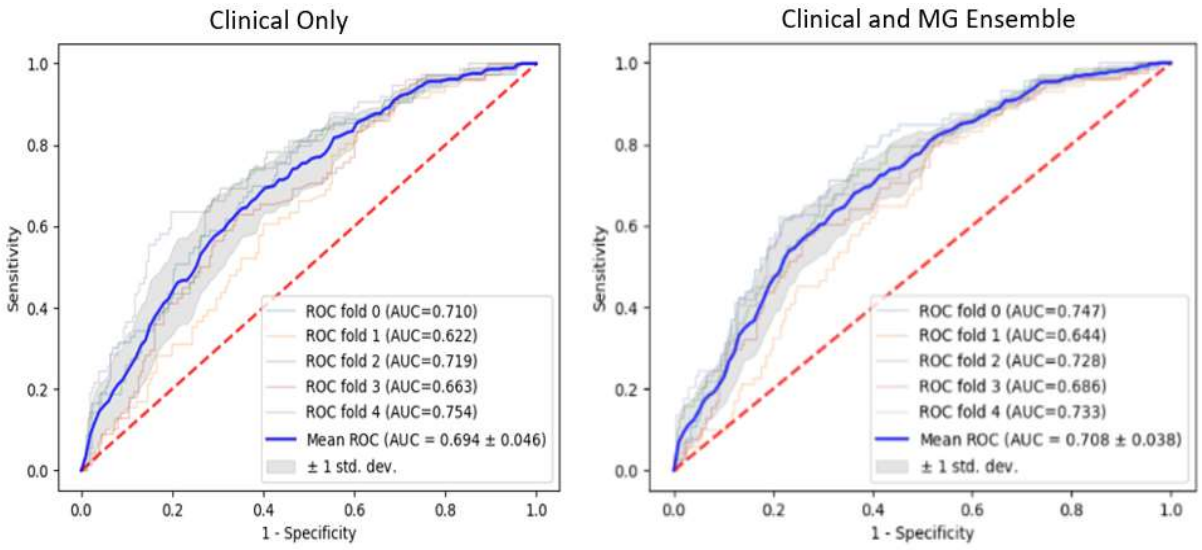

Figure 4. The 5-folds cross validation ROC curves and AUC. Left: Model for clinical data alone with an average AUC across all folds of 0.69. Right: Ensemble of clinical and MG models with an average AUC score from all folds of 0.71.

6. DISCUSSION

In this paper we show an ability to predict pCR using multi modal algorithms that include features extracted from MG images and clinical data prior to neoadjuvant chemotherapy treatment. In MG scans, the features are extracted both from the tumor area and from the peritumoral area. In the clinical data, our results show the importance of features such as estrogen, progesterone, HER2 and KI67 to the classification task, and this result coincide with results from previous work[12]. These features are important indicators for the success of NAC therapy as we can see their significance reflected in the p-value.
Our data cohort is small and so training a deep learning network on that data is a challenge. We examined the approach of combining a pre-trained DL model with texture features extraction and show that useful information for the prediction of our task may exist in the MG data.

MR image utilization was less efficient with Curie data due to high variance on MR image modalities. For MRI, there is no standardized protocol for scan acquisition, and thus one scan can have multiple scan intervals post contrast agent injection while others have only one. Scan quality can also vary according to the scanner magnet (1.5T or 3T) and different resolutions expressed by the mm to pixel value. The most significant challenges are connected to high variance of image resolution and voxel size as well as high variance on image contrast dynamics. Images were collected from multiple hospital sites and multiple MR devices within these sites which lead to significant differences on image noise. These lead easily for example to situations that ML algorithms start detecting sub cohorts of patients based on site or machine used to measure them. Next steps with MRI modalities are to find a robust method for segmenting the breast tissue out from other volume within the current challenging data set. This will enable to normalize the features from tumor segmentation over the whole patient cohort.

In our problem setting high sensitivity is crucial as we want all patients that achieved pCR to also be correctly classified by our model. Within this high sensitivity, the resulted specificity in our ensemble model show that for 22% of the patients, the model could correctly predict in advance that they won't achieve pCR enabling them to reassess in advance this treatment.

7. CONCLUSION

This study introduces a radiomics system which can work on clusters on-premise or in-the-cloud, and display results in a visualized manner. The system predicts NACT pathologic complete response by analyzing patient's mammograms and clinical information using a multimodal ensemble model. Achieving complete response in this treatment is correlated with improved overall survival compared with those experiencing a partial or no response at all. For Curie cohort of 1383 patients, our final model achieves 0.708 AUC. The model can predict in advance most of the patients that will achieve pCR (sensitivity 0.954) and some of the patients that will fail to achieve that (specificity 0.222), enabling them to reassess in advance this treatment.

Future work may add US imaging analysis that already exist in Curie NACT cohort. For MRI, we plan to utilize deep learning methods that will be more robust to deal with the high variance in the MRI dataset. We also seek to use bigger cohorts and from additional sites to increase our training data and have better generalization, towards large scale validation of our models.

ACKNOWLEDGEMENT

We thank Johan Archinard from Institut Curie for his dedicated continuous support with the IT infrastructure for this pilot. We thank Prof. Fabien Reyal, Dr. Caroline Malhaire and Dr. Anne Sophie Hamy-Petit from Institut Curie for defining the clinical use case, share their experience and help understanding the data.
Research reported in this publication was partially supported by European Union's Horizon 2020 research and innovation program under grant agreement No 780495. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of data appearing therein.

REFERENCES

[1] Teshome, M. and Hunt, K. K., "Neoadjuvant therapy in the treatment of breast cancer", Surgical oncology clinics of North America, 23(3), 505–523 (2014).

[2] Cortazar, P. et al., "Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis", The Lancet 384, 164–72 (2014).

[3] Von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J, Jackisch C, Kaufmann M, Konecny GE, Denkert C, Nekljudova V, Mehta K, Loibl S., "Definition and Impact of Pathologic Complete Response on Prognosis After Neoadjuvant Chemotherapy in Various Intrinsic Breast Cancer Subtypes", Journal of Clinical Oncology (2012).

[4] Gillies, R. J., Kinahan, P. E., Hricak, H., "Radiomics: images are more than pictures, they are data", Radiology, 278(2), 563-577 (2015).

[5] Big Data for Medical Analytics, BigMedilytics EU Project, https://www.bigmedilytics.eu/

[6] Parekh, Vishwa, Jacobs, Michael., "MPRAD: A Multiparametric Radiomics Framework", arXiv:1809.09973 (2018).

[7] Scheel, J. R., Kim, E., Partridge, S. C., Lehman, C. D., Rosen, M. A., Bernreuter, W. K., & Polin, S. M., "MRI, clinical examination, and mammography for preoperative assessment of residual disease and pathologic complete response after neoadjuvant chemotherapy for breast cancer", ACRIN 6657 Trial. American Journal of Roentgenology, 1376-1385. (2018).

[8] Savaridas, S. L., Sim, Y. T., Vinnicombe, S. J., Purdie, C. A., Thompson, A. M., & Evans, A., "Are baseline ultrasound and mammographic features associated with rates of pathological completes response in patients receiving neoadjuvant chemotherapy for breast cancer?", Cancer Imaging, 19: 67, (2019).

[9] Ring AE, Smith IE, Ashley S, Fulford LG, Lakhani SR., "Pathological complete response and prognosis in patients receiving neoadjuvant chemotherapy for early breast cancer", British Journal of Cancer, 91(12):2012–7 (2004).

[10] Jun-jie Li, Canming Chen, Yajia Gu, Genhong Di, Jiong Wu, Guangyu Liu, ZhiMin Shao, "The role of mammographic calcification in the neoadjuvant therapy of breast cancer imaging evaluation", PLoS One, 9(2), (2014).

[11] Lee, H., Lee, D. E., Park, S., Kim, T. S., Jung, S. Y., Lee, S., ... & Lee, K. S., "Predicting response to neoadjuvant chemotherapy in patients with breast cancer: combined statistical modeling using clinicopathological factors and FDG PET/CT texture parameters", Clinical Nuclear Medicine, 44(1), 21-29, (2019).

[12] Fasching, P.H., "Ki67, chemotherapy response, and prognosis in breast cancer patients receiving neoadjuvant treatment", BMC Cancer, 486 (2011).

[13] Ravichandran, K., Braman, N., Janowczyk, A. and Madabhushi, A., "A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI", proceedings of SPIE Medical Imaging, (2018).

[14] Chen, T., Guestrin, C., "Xgboost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. (2016).

[15] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence. (2017)

[16] Le, T. L. T., Thome, N., Bernard, S., Bismuth, V., & Patoureaux, F. Multitask classification and segmentation for cancer diagnosis in mammography. arXiv preprint arXiv:1909.05397. (2019).

[17] HL7 Fast Healthcare Interoperability Resources (FHIR), https://www.hl7.org/fhir/overview.html